UNSUPERVISED LEARNING: WHOLESALE SPENDING ANAYLSIS

1 INTRODUCTION AND RESEARCH QUESTIONS

This assignment performs unsupervised learning methods on the wholesale data set to discover unknown patterns and homogeneous subgroups within the observations. Within unsupervised learning there are only a set of X features measured on n observations. Therefore, the task is not to formulate a prediction as we do not have an associated response variable Y. [1] Instead, this task aims to uncover meaning inside the homogeneous subgroups – or clusters – unsupervised learning finds. In this assignment, the X features, correspond to wholesale product categories: fresh, milk, grocery, frozen, detergents and paper and delicatessen. The n observations correspond to values of annual spending in monetary units (m.u.) by the clients of the wholesale distributor. The main objective of using unsupervised learning on this data is to understand of how principal component analysis (PCA) and clustering techniques find subgroups in the data. Notably, PCA reduces dimensionality by forming new combined variables from a dataset. PCA is an important as the loss of dimensions, but retention of variability affects cluster patterns in the data. For this reason, the primary question asked in this assignment is: *does principal component analysis improve cluster dissimilarity, density, and separation?* To answer this question, this assignment first explores outliers and relationships between variables in the initial data set. Secondly, missing observations are dealt with through imputation. The imputed data set is then used in PCA to produce a dimensionally reduced data set. Lastly, this assignment compares the outcome of clustering methods on the dimensionally reduced data to the original data. This comparison should highlight differences in cluster density and separation depending on whether the data has undergone dimension reduction. Therefore, the utility of this assignment is to discover if loosing redundant information helps towards making more distinct clusters.

2 EXPLORATIONS OF THE DATA

This section seeks to gather an understanding of the data set by observing initial patterns in the data. This includes observing outliers, relationships between variables and missing data values.

2.1 OUTLIERS

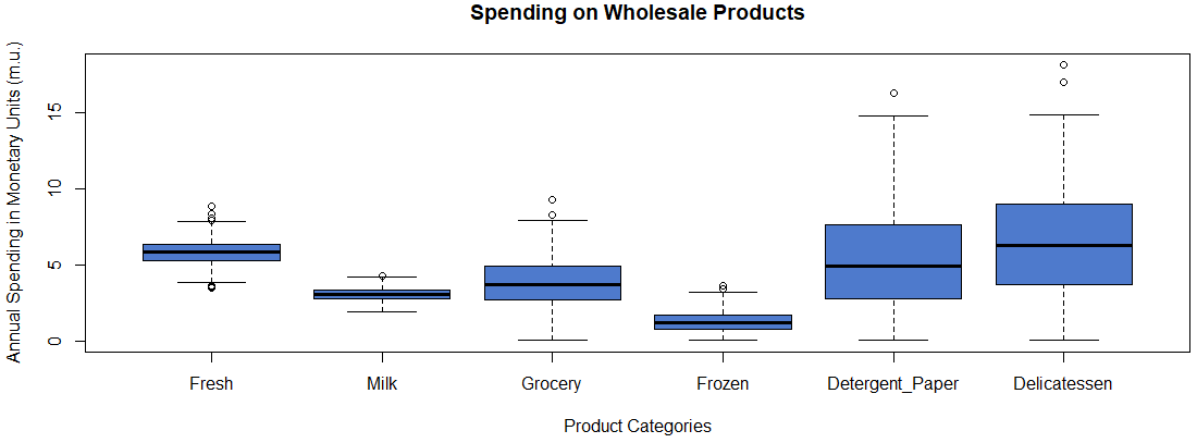


Figure 1. Box plots for wholesale product data showing outliers (circles) for each variable.

The box plot Fig.1 shows product categories on the x-axis plotted against annual spending in monetary units on the y-axis. Notably, fresh products have the most outlier values in both lower and higher values than expected. This suggests customers sometimes spend either higher or lower amounts than expected on fresh products. By contrast, customers sometimes spend higher than usual on products other than

[1] James, (2020), 497.

fresh – but never spend lower than usual on products other than fresh. Table 1 provides further explanation for this outlier spending behaviour. The outlier values are listed with their corresponding row name (ID) and variable. The ID is the row index in the original dataset since ID does not exist as a variable.
The box plot shows the highest spending is on delicatessen products. Delicatessen has the highest median spending annually, and the highest interquartile range. Moreover, the highest spending amounts were for delicatessen products, which are outliers at 16.94 (m.u) and18.090 (m.u). The least overall spending is on frozen products. This is evident as the median spending on frozen products is low 1.170 (m.u). Additionally, frozen products have the lowest outlier values at 3.63 and 3.43 (m.u).

Table 1. List of outliers identified in the data set with their corresponding row name (ID) and variable.

ID	Product	Value (m.u)
336	Fresh	3.50
439	Fresh	3.55
307	Fresh	3.63
164	Fresh	8.84
427	Fresh	8.36
371	Fresh	8.32
16	Fresh	8.06
196	Fresh	8.06
125	Fresh	8.03
265	Fresh	7.90
176	Milk	4.27
164	Grocery	9.24
360	Grocery	8.26
164	Frozen	3.63
360	Frozen	3.43
190	Det.P	16.26
167	Delica.	18.09
232	Delica.	16.94

2.2 RELATIONSHIPS BETWEEN PRODUCTS

The scatter plot matrix Fig 2. shows associations between variables. For example, strong linear associations between variables show as compact observations which monotonically increase or decrease. By contrast, weaker associations reveal random or dispersed observations on the scatter plots. The correlation coefficients are shown in the top right section of the matrix. Strong relationships have coefficients close to 1, while weak relationships have coefficients close to 0. Fresh is highly related to grocery and frozen products. They share a strong positive linear relationship with each other and have high coefficients: 0.87 and 0.80. This implies that these three variables share redundant information. Therefore, PCA can be used to remove this redundancy. Milk has a negative relationship to grocery and frozen variables, as shown by the scatter plots and the coefficient values: -0.42 and -0.36. This shows there is some negative association between spending on milk and spending on grocery and frozen products, however the association is not especially strong. Detergent and paper and delicatessen products seem to have little linear relationship to any variables. Instead, these variables generally cluster into one group. Fresh and delicatessen is the exception, this pattern shows a main cluster and a sparse group of observations at higher values of fresh. Therefore, as there may be some clustering patterns in the data, clustering methods will be used in the data analysis – section 4.

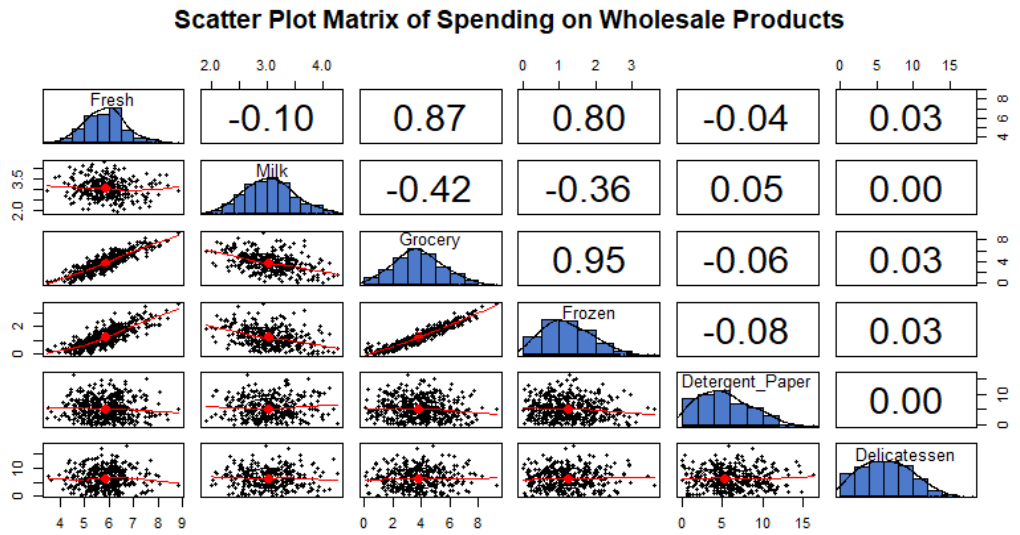


Figure 2. Scatter plot and correlation coefficient matrix showing strengths of linear relationships

2.3 MISSING DATA

Before proceeding to model the data, it is important to explore the pattern of missing values in the data. In this section the package 'VIM' is used to visualise and summarise missing data. Table 2 shows overall there are 270 complete cases. The second row indicates that there are 78 cases with only one missing value, which is on Milk. In the third row there are 40 cases with only one missing value, in delicatessen. Only two cases have missing values, which are on delicatessen and milk.

The last row of table 2 shows how many values are missing from each variable (column). Fresh, grocery, frozen and detergents and paper have no missing values. By contrast, 92 observations are missing in delicatessen, and 130 are missing in milk. Therefore, between delicatessen and milk there are 222 missing observations. Fig 3. Visualises the missingness pattern, where blue represents observed data and red represents missing data. 61.4% of observations are not missing any information, 17.7% are missing milk values, while 9.1% of observations are missing delicatessen values. The last 11.8% of observation are missing both milk and delicatessen values.

Table 2. Missing Data Cases

Complete Cases	Fresh	Grocery	Frozen	Det.P	Delica.	Milk	Missing Cases
270	1	1	1	1	1	1	0
78	1	1	1	1	1	0	1
40	1	1	1	1	0	1	1
52	1	1	1	1	0	0	2
	0	0	0	0	92	130	222

Fig 4. helps visualise the missing data patterns from milk and delicatessen in a margin plot. The blue box plots summarize the distribution of observed data given the other variable is observed. The red box plots summarize the distribution of observed data given the other variable is missing. For example, the red box plot on the x-axis shows lower values of delicatessen observations are missing. There are 92 observations in which delicatessen is missing and 52 instances in which both milk and delicatessen are missing. As only values are missing from milk and delicatessen products, the missing values are most

likely missing at random (MAR). MAR means the missingness is not completely random, but the propensity of missingness depends on the observed data, not the missing data.[2] This suggests there is reason why milk and delicatessen products are missing. On the other hand, as there is no apparent reason why milk and delicatessen products have missing values, this pattern may be a coincidence. Therefore, it is possible the data is missing completely at random (MCAR).

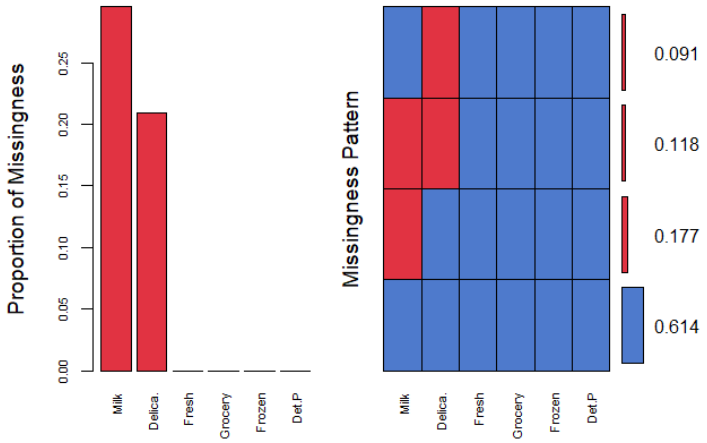


Figure 3. Missing data plots showing proportion of missingness and missingness pattern

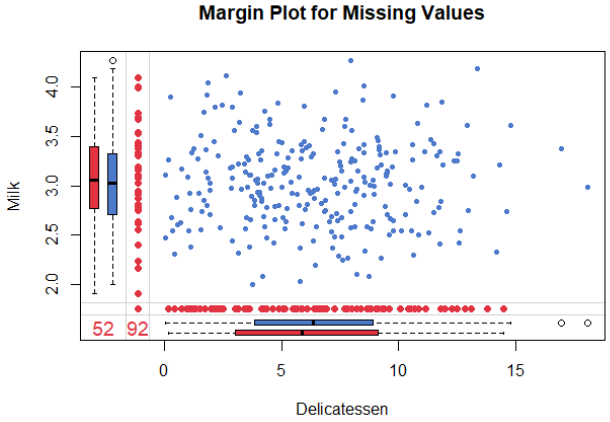


Figure 4. Missing data margin plot summarising the distribution of observed data given the other variable is observed.

3 METHODOLOGIES

The methodology for analysing the wholesale data follows four steps. Firstly, the original data set is explored (in section 2). Secondly the data is pre-processed for further analysis. This involves dealing with missing data using imputation. Imputation replaces missing observations with realistic observations. Importantly, this step is crucial as there is a large fraction of missing observations in the data – 222 out of 440. Consequently, removing missing values would be wasteful. Moreover, unsupervised learning methods cannot handle missing values. Therefore, without dealing with missing observations, data analysis could not be performed. Section 3.1 further explains the method of imputation. The third stage in this methodology is to reduce the dimensions of the data. This stage is

[2] 'Getting Started with Multiple Imputation in R'. University of Virginia Library. Website.

pinnacle to the understanding the question proposed in the introduction: *how does dimension reduction affect cluster dissimilarity, density, and separation.* For dimension reduction, principal component analysis is used. PCA is an important method to analyse correlations inside a dataset but also to form new combined variables. Therefore, the new combined variables created in PCA most likely form new patterns within the data which affect clustering results. The fourth stage in this methodology implements clustering methods on the data. Namely, hierarchical clustering and k-medoid clustering. These two methods are used since between them, they provide a wide scope of clustering analysis of the data. They implement different algorithms to discover clusters and use two different visualisations. At this stage clustering is used to find homogeneous subgroups among the observations.[3] The results of clustering will be validated using silhouette analysis. This validation will provide a conclusive assessment of how dimension reduction affects the validity of clusters, in addition to how suitable clustering is for the wholesale data.

3.1 IMPUTATION

To deal with missing values, this assignment uses multiple imputation techniques but only extracts one data set, like in single or regression imputation. Extracting one of the imputed data sets is necessary to carry out principal component analysis on missing data. Multiple imputation is more beneficial for this unsupervised learning task than single imputation as there is no requirement to specify a model before imputing the data. This means no estimates about variable relationships are introduced before the missing data is imputed. Moreover, multiple imputation assumes the data is MAR.[4] This aligns with our assumption that the data is most likely missing not at random since the only missing values come from milk and delicatessen products.

To impute the data, the mice package is used. Mice imputes each missing value with a plausible value *m* number of times. In this assignment m is set to 10, therefore 10 values were imputed for each missing observation. This process is completed when all missing values are imputed *m* times, and the data set is completed. Any one of these imputed data sets can be used for dimension reduction in PCA.

Density plots in Fig 5. show the results of the first imputed data set using different methods of generating imputed data. Predictive mean matching (pmm) involves selecting a data point from the original, non-missing data which has a predicted value close to the predicted value of the missing sample. By contrast, weighted predictive mean matching (midastouch) uses a distance aided selection of donors to predict the missing data.[5] The random forest (rf) method imputes the missing data by generating random forests.[6] The blue line in the density plots show the observed value, while the red line shows the predicted value. Therefore, the most effective method is weighted predictive mean matching. This is because this method shows the most agreement between the observed and imputed values for both variables, meaning the predicted values are indeed plausible. Therefore, the first imputed data set generated through weighted predictive mean matching is used throughout the assignment.

[3] James, (2020), 517

[4] 'Getting Started with Multiple Imputation in R'. University of Virginia Library. Website. The other option is single imputation. However, single imputation does not pool the imputed values together to provide an average imputed value.

[5] 'mice.impute.midastouch: Predictive Mean Matching with Distance aided selection of donors'. Rdocumentation. Website.

[6] 'mice.impute.rf: Imputation by random forests'. Rdocumentation. Website.

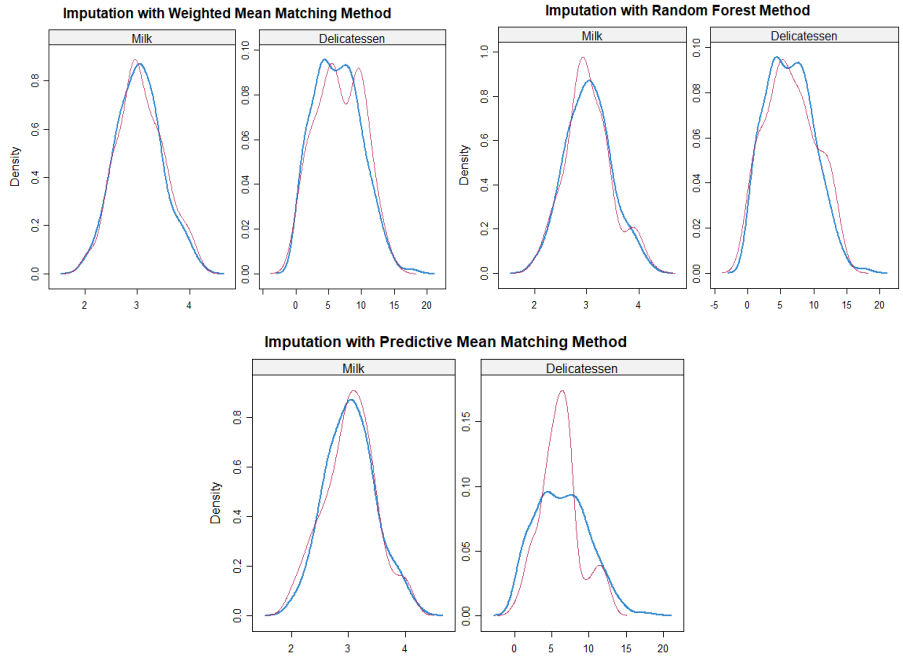


Figure 5. Density plots showing the difference between observed values (blue) and imputed values (red).

4 DATA ANALYSIS, RESULTS, AND INTERPRETATION

This section performs the unsupervised learning methods PCA and clustering to answer the primary question: *does principal component analysis improves cluster dissimilarity, density, and separation.* Clustering will be performed on the original data to create the baseline used to compare against the dimensionally reduced data set generated by PCA.

4.1 PRINCIPAL COMPONENT ANAYLISIS

In this section we explore how to reduce the dimensions of the data set through principal component analysis (PCA). This is useful as PCA explains most of the data original data set in fewer dimensions.[7] The dimensionally reduced data set can be then utilised in further modelling processes such as clustering. PCA is an unsupervised approach, since it involves only a set of features, and no associated response.[8] In other words, PCA provides a way of visualising n observations with measurements on a set of p features. The concept is that each of the n observations lives in p-dimensional space, but not all these dimensions are equally informative. For this reason, PCA reduces the dimensions by finding components which have linear combinations of the p features. For example, the first principal component of the features is the normalised linear combination of the features which has the largest variance.[9] (1)

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p \qquad (1)$$

To determine the optimum number of dimensions to retain, we must look at the proportion of the variance that is explained by each of the dimensions. Table 3 shows each dimension and the

[7] James, (2020), 498.
[8] James, (2020), 499.
[9] James, (2020), 499 – 500.

corresponding eigen value and variance. In addition, the cumulative variance for each dimension is shown in the last column. Note the first principal component has the highest variance. Based on table 3, either three or four dimensions should be kept. This is because three or four dimensions cumulatively retain at least 80 – 90% of the variability in the data set while still reducing the dimensions. Moreover, there is a sudden drop in variance after four dimensions. This means dimensions five and six can be taken out of the data set as they contribute only a small percentage of variance.

Table 3. Dimensions, Eigen Values and Variance for the Wholesale Data

Dimensions	Eigen Value	Variance	Cumvariance
1	2.92	48.62	48.63
2	1.03	17.24	65.87
3	1.01	16.77	82.64
4	0.85	14.09	96.74
5	0.17	2.78	99.51
6	0.03	0.49	100

The proportion of variance retained by each dimension can be plotted using a scree plot. (Fig 6.) Importantly, the data must be scaled to have a standard deviation of one. Fig 7. shows when the data is not scaled the first two components alone explain 84.8 % of variance. This means that nearly all weight is placed on the first and second loading vector.[10] Moreover, the proportion of variance explained in the unscaled data does not reflect the eigen values corresponding to each component. The proportion of variance explained by a component is equal to the eigenvalue of a component divided by the sum of eigen values. For example, using the first eigen value, the variance explained by the first component is calculated by the following sum: (2)

$$2.91752930 \cdot 100 / 6 = 48.625 \qquad (2)$$

This answer (48.6) corresponds to the scaled data.

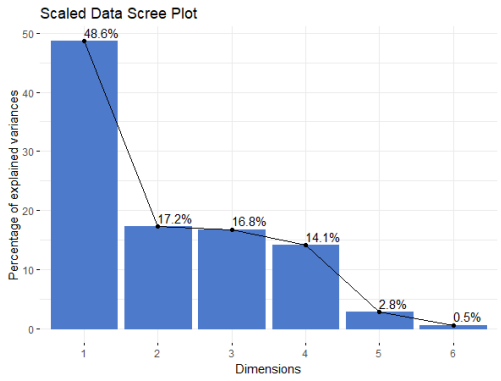


Figure 6. Scree plot depicting the proportion of variance explained by each of the six principal components in the scaled wholesale data.

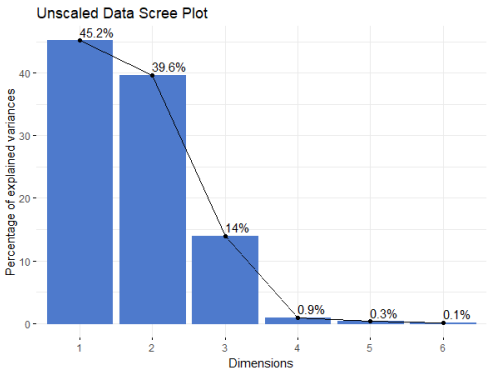


Figure 7. Scree plot depicting the proportion of variance explained by each of the six principal components in the unscaled wholesale data.

Analysing how variable information contributes to the dimensions helps in determining whether to use three or four dimensions. Fig 8 & 9. show the contributions of each variable to the dimensions in the data set. As there are 6 variables in the data set, any variable with height up to 17 and above is considered

[10] James (2020), 507.

to have contributed significantly to the component (100/6 ≈ 17). In this case there is a sudden drop in the contributions of detergents and paper and milk when only three dimensions are retained. By contrast, in dimension four each variable contributes either the significant amount of information or very nearly the significant amount of information. Therefore, four dimensions reduce the dimensions of the data and keep most variability and information. For this reason, PC1, PC2, PC3 and PC4 are the components retained in the reduced dimension data set.[11] This reduced data set will be subsequently analysed using clustering methods in section 4.2.

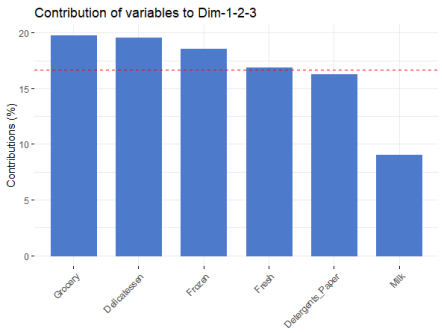


Figure 8. Variable contributions to the first three dimensions of the dataset.

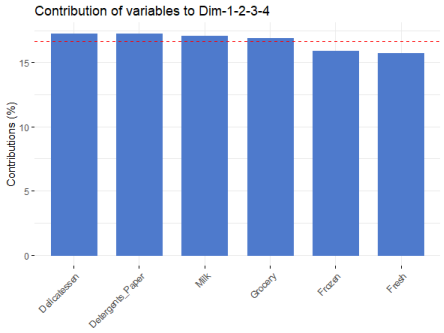


Figure 9. Variable contributions to the four dimensions of the dataset.

Meaning from the data can be seen in the results of principal component loading vectors. Fig 10. shows the biplot of the first and second principal component loading vectors, while Fig 11. is the biplot of the third and fourth. The loadings are also given in biplot Table 4. The biplot in Fig 11. and table show the third loading vector places most of its weight on delicatessen products. However, the fourth loading vector places much less weight on delicatessen. For example, the loading for delicatessen on the third component is 0.98, and its loading on the fourth principal component 0.19 (the word Delicatessen is centred at the point (0.98, 0.19)). Fig 11. suggests the product delicatessen, with a positive score on the third component, has a higher spending rate. By contrast the product milk, with a negative score on the third component, has a lower spending rate

Grocery, frozen, fresh and milk products seem to span a negative vector in the fourth principal component. Hence the fourth component accounts for the correlation in spending between these products.[12] Fig 10. shows that there is also an association between fresh, grocery and frozen, as they are grouped close away from other products in dimension 1. Interestingly, in both biplots detergent and paper is further away from other products. This suggests spending on detergents and paper is less correlated to spending on all other products.

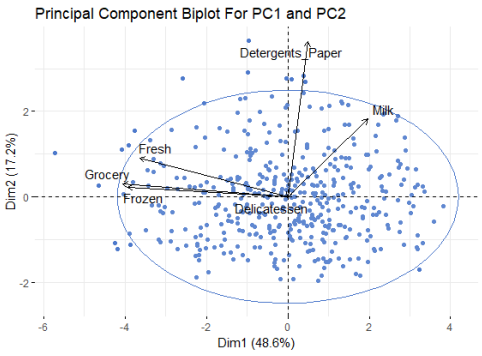


Figure 10. Biplot of the first two principal components for the wholesale data.

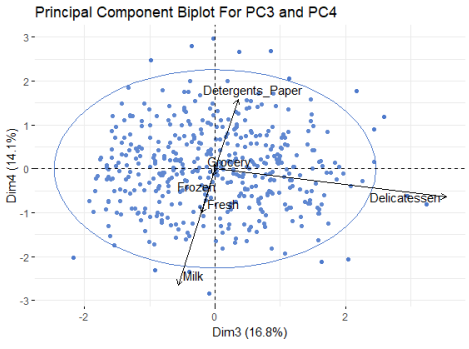


Figure 11. Biplot of the third and fourth principal components for the wholesale data.

[11] See appendix 1 for the contribution of variables to individual components.

[12] Note, this correlation can be initially seen in the scatter plot matrix.

Table 4. The Principal Component Loading Vectors for the Wholesale Data.

Variable	Principal Component Loading Vectors			
	PC1	PC2	PC3	PC4
Fresh	0.52	-0.21	-0.06	0.30
Milk	-0.28	-0.44	-0.16	0.80
Grocery	0.58	-0.07	-0.004	0.01
Froz	0.56	-0.05	-0.01	0.04
Det.P	-0.07	-0.87	0.09	-0.48
Delica.	0.001	0.005	0.98	0.19

4.2 CLUSTERING

This section performs hierarchical and k-medoid clustering on the dimensionally reduced data and the original data set. These two clustering methods provide a broad analysis of the data as they utilise different algorithms to create the clusters and present different visualisations of the results. Clustering methods will help uncover meaning in the data by separating the data into subgroups or clusters of homogenous observations. Here it must be noted that for consistency in methodology all clustering methods use Euclidean distance measures.

4.2.1 HIERARCHICAL CLUSTERING

Hierarchal clustering is the first clustering approach used in this assignment. This method is chosen as it does not require a pre-specified number of clusters, since the number of clusters present can be determined from hierarchical clustering's tree-based visualisation. Moreover, rather than performing hierarchical clustering on the entire data set or just two principal components at a time, all four principal components in the dimensionally reduced data set can be clustered.[13] This will provide a wide scope of clustering analysis when comparing clustering methods in section 5.

In hierarchical clustering, a hierarchy of partition is created, at each level two or more groups are merged to the ones of the previous level. There are two types of hierarchal clustering, this assignment uses the bottom-up or agglomerative clustering. Agglomerative clustering first makes each observation its own cluster. Until there is only one cluster left, agglomerative clustering merges groups which have the smallest dissimilarity – or smallest distance. The dissimilarity between a pair of observations must be extended to a pair of groups of observations. This extension is achieved by developing the notion of linkage. Linkage defines the dissimilarity between two groups of observations. The four most common types of linkage are: complete, average, single, and centroid.[14]

To determine which method of linkage is best, the map_dbl() function from tidyverse is used to obtain the agglomerative coefficients for the dimensionally reduced data. (Table 5.) The highest ac value from the linkage results corresponds to the method that maximises dissimilarity between clusters. Therefore, in this assignment complete linkage at 93% is the most suitable method as it makes the clusters distinct.[15]

Table 5. Ac Values for Linkage Methods

Average	Single	Complete
0.84	0.71	0.93

[13] James, (2020), 547.

[14] James, (2020), 527.

[15] James, (2020), 545. To see the graphic results of different linkage methods, see appendix 2. Note that complete linkage yields the most balanced and distinctive clusters. Average linkage method creates two unbalanced clusters. By contrast the single linkage method causes observations to trail into large clusters, onto which individual observations attach one-by-one.

The results of hierarchical clustering using complete linkage are shown by the dendrograms in Fig 12 & 13. Fig 12. shows these results based on the dimensionally reduced data (the first four principal components), while Fig 13. shows the results based on the original data set. Clusters in the dendrograms can be read by looking at where the horizontal bars, or parent nodes, spilt. The distinct sets of observations beneath the spilt are interpreted as clusters. In these dendrograms, the main parent node separates into two distinct branches, revealing two clusters in the data. In the lower part of the dendrogram nodes fused to parent nodes correspond to observations that are like each other. In contrast, the nodes which fuse together towards the top of the tree can be different to each other. Importantly, the split is less equal in the dimensionally reduced data set. (Table 6) Therefore, it is likely there is more dissimilarity between the dimensionally reduced clusters. By contrast, as the clusters are roughly in equal in size when the data is not dimensionally reduced, observations between these clusters may be similar.

Table 6. Observations in Each Cluster

Clusters	Original Data	Four Principal Components
1	232	297
2	208	143

Fig 12. & 13. also show dimension reduction increases dissimilarity between the clusters. There is a greater difference between the heights of the clusters in the dimensionally reduced data. Obtaining the height of the first two branches reveals the orange cluster has a height of 7.6, while the blue cluster's height is 6.9. By contrast, without dimension reduction, the heights are more similar. The blue cluster has a height of 6.7 and the orange cluster is 7. Importantly, the height of the cut to the dendrogram, controls the number of clusters created – effectively serving the same role as a medoid or centroid.[16] Therefore, the two clusters are more dissimilar from each other after dimensional reduction. Overall, the results of hierarchal clustering differ depending on the dimension input. In this case, dimension reduction improves cluster dissimilarity. For this reason, we can argue dimensional reduction denoises the data.[17]

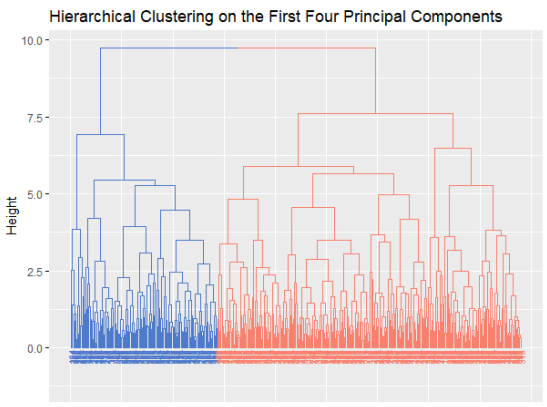


Figure 12. Dendrogram of hierarchal clustering results on dimensionally reduced data.

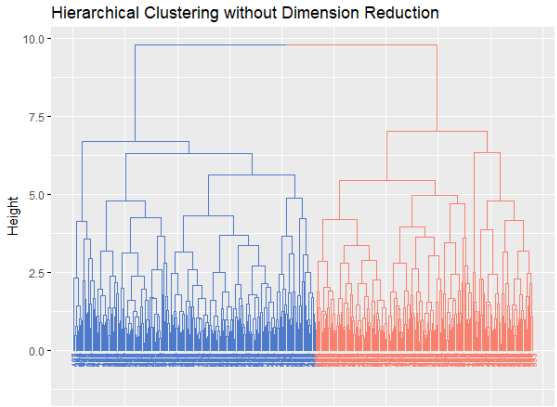


Figure 13. Dendrogram of hierarchal clustering results on the original data.

[16] James, (2020), 524.
[17] James (2020), 547.

4.2.2 K – MEDOID CLUSTERS

The next clustering method used in this assignment is k-medoid clustering. K-medoid clustering is useful as it is robust to outliers. Although the proportion of outliers is low (18 out of 440), outliers may still affect the clusters since the wholesale data contains at least one outlier in each product category. K-medoid clustering is less sensitive to outliers than other clustering methods as it uses medoids for the centre of clusters, which are real data points. This contrasts k-means clustering which uses centroids – the average of the members of the cluster to form the cluster centre.

Unlike hierarchal clustering, the number of clusters must first be specified in k-medoid clustering. Therefore, before beginning the k-medoid clustering, the optimal number of clusters is determined through silhouette analysis. Silhouette analysis computes different values of k clusters, then the average clusters silhouette is computed according to the number of clusters. The average silhouette measures the quality of clustering. A high average silhouette width indicates a good clustering. The optimal number of k clusters is the one that maximizes the average silhouette over a range of possible values for k.[18] Here Fig 14. & 15 and show the optimum number of clusters for the dimensionally reduced data set and the data set without dimensional reduction.

To fit the k-medoid clusters, the first and fourth principal component score vectors are used. This is because, as previously noted in the scree plot Fig 6. the first component has the highest variance. Moreover, components 1 and 4 collectively share significant contributions from most variables. (Fig 16. & 17.)

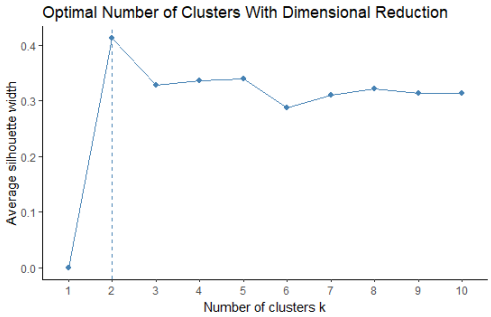


Figure 14. Silhouette analysis showing the optimal number of clusters for the dimensionally reduced data set.

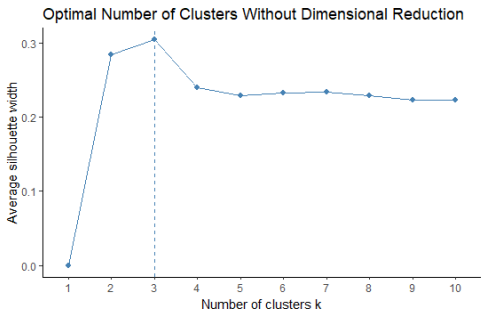


Figure 15. Silhouette analysis showing the optimal number of clusters for the original data.

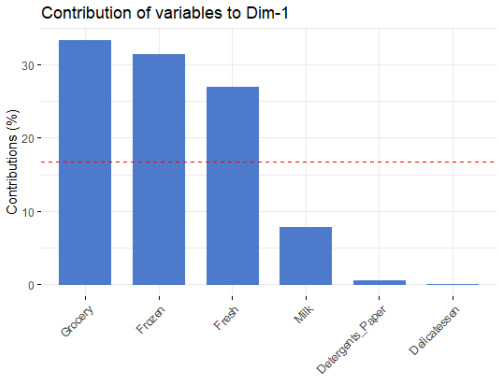


Figure 16. Variable contributions to the first dimension of the dataset.

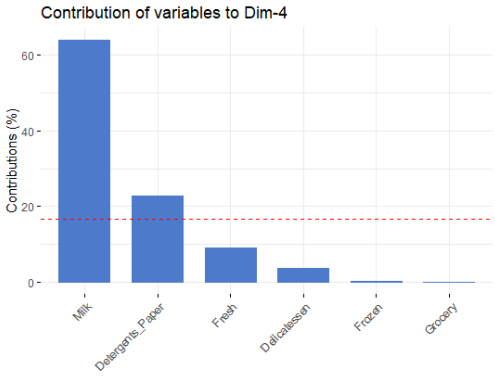


Figure 17. Variable contributions to the fourth dimension of the dataset.

[18] Kassambara, (2017), 52.

Now, the k-medoid clusters are fitted with the PAM (Partitioning Around Medoids) algorithm. PAM is used as the data is not that large, we want to use all the data and not just a sample space.[19] Tables 7 & 8. show the medoids for each cluster and their corresponding component names. Note that in the dimensionally reduced data set variables names are lost, and instead the medoids for principal components 1 and 4 are given. Importantly, dimension reduction captures more variability in each cluster with fewer components. Without dimension reduction the medoids can be distinctly different or very much like each other. For example, the medoid for delicatessen in cluster 1 is 9.53, while the medoid for cluster 2 is 3.65. On the other hand, the medoids for Milk are very similar for each cluster. By contrast the medoids for PC1 and PC4 consistently capture variability. (Table 8.)

Table 7. Medoids for Original Data

Cluster	Fresh	Milk	Grocery	Frozen	Det.P	Delica.
1	5.89	3.16	4.28	0.93	3.81	9.53
2	5.53	3.11	3.35	0.98	3.59	3.65
3	6.37	3.09	3.67	1.11	9.59	5.74

Table 8. Medoids for Dimensionally Reduced Data

Cluster	PC1	PC4
1	0.99	-0.03
2	-1.36	0.15

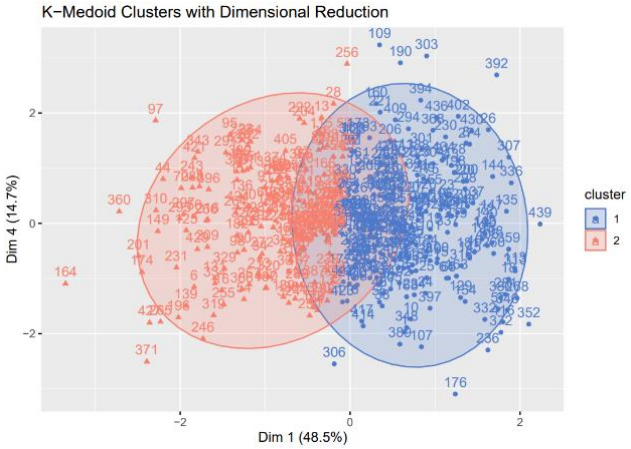


Figure 18. Plot of k-medoid clustering on the dimensionally reduced data.

Fig 18. plots the results of clustering on the dimensionally reduced data set. Importantly, there is some overlap between the two clusters. Moreover, there are multiple observations outside of the cluster. Ideally, observations in the cluster should be closer to members within that cluster rather than members in another cluster. Therefore, we would hope to see dense clusters that are well separated in the visualisation. However, this is not the case. As dimension reduction did not create separated clusters, it is likely the wholesale data may not be suitable for clustering. Nevertheless, comparing the clusters created from the original data shows dimension reduction greatly improves the cluster results. (Fig 19.) Visualising all dimensions in Fig 19. shows nearly all clusters completely overlap. In Fig 19. only dimensions 2 and 3 resemble unique clusters. However, these clusters still overlap more than the clusters in Fig 18. Overall, this highlights how dimension reduction improves cluster density, dissimilarity, and separation.

[19] Kassambara, (2017), 56-57.

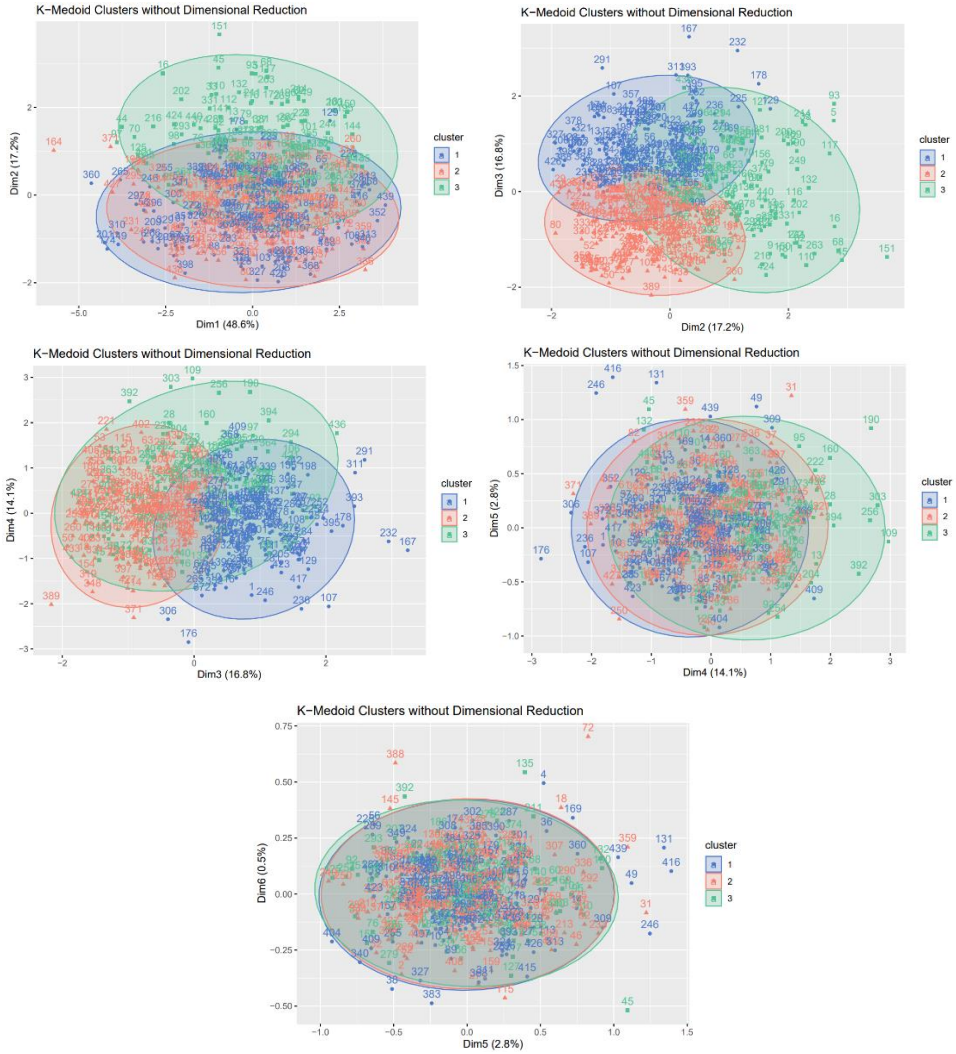


Figure 19. Plots of k-medoid clustering on the original data set.

4.2.3 CLUSTER VALIDITY

This section will compare the validity of clustering on the dimensionally reduced data set against the original data set using silhouette analysis. Silhouette analysis measures the compactness and separation of cluster to assess their validity. Silhouette is based on the difference between the average distance of points in the closest cluster and to points in the same cluster. The optimum cluster method can be determined based on the average silhouette width which lies in the interval [-1, +1]. A value close to +1 is ideal, indicating an observation is close to observations in its own cluster and far from observations in other clusters. A value close to 0 indicates that an observation is close to the boundary between clusters. A value close to – 1 indicates an observation is closer to observations from another cluster than its own and is potentially miss clustered.

Table 9. shows the cluster size and average silhouette width for the dimensionally reduced data set and the original data. Fig 20. & 21. visualise the average silhouette width for each cluster and data set. For the original data set the silhouette average widths are close to 0, ranging from 0.11 to 0.15 with an average width of 0.13. The silhouette widths for the dimensionally reduced data set are a little further away from 0, ranging from 0.24 to 0.27 with an average width of 0.26. As the scores are low this means the observations within clusters in both data sets are close to the boundary of other clusters. This pattern is very apparent in k-medoid clustering Fig 18. Therefore, validation shows neither data set produces

strongly dissimilar and separated clusters. Overall, this indicates clustering is not a suitable method for the wholesale data set.

However, it must be noted that silhouette widths are higher for the dimensionally reduced data set. This shows dimension reduction increased the validity of the clustering methods. Therefore, to answer the primary question, dimension reduction does indeed improve cluster analysis by producing more dissimilar, dense, and separated clusters.

Table 9. Cluster Size and Average Silhouette Width

	Reduced Dimensions – PC 1:4		Original Data Set	
Cluster	Cluster Size	Ave.sil. Width	Cluster Size	Ave.sil. Width
1	239	0.27	156	0.12
2	201	0.24	173	0.15
3	-	-	111	0.11

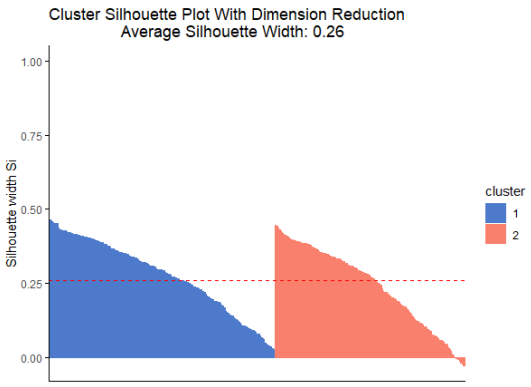


Figure 20. Cluster Silhouette plot on the dimensionally reduced data with 2 clusters.

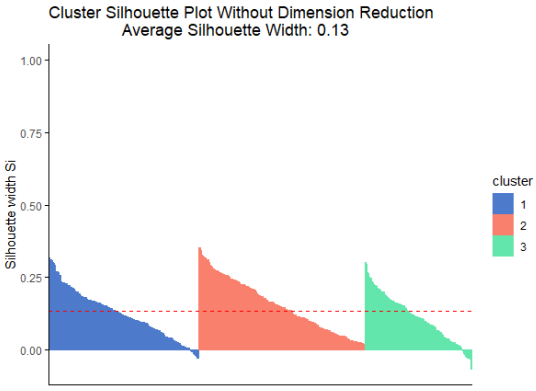


Figure 21. Cluster Silhouette plot on the original data with 3 clusters.

5 DISCUSSION AND CONCLUSION

This section discusses the utility of PCA and clustering on the wholesale data set. Then conclusions are made about what meaning unsupervised methods have uncovered in the data.

The results obtained from hierarchal clustering and k-medoid clustering consistently provide evidence that PCA improves cluster dissimilarity, density, and separation. This is because the clusters created from dimensionally reduced data set are more separated, dense, and dissimilar to clusters created from the original data set. Moreover, the silhouette width average is higher after PCA is applied. This means reducing dimensions through PCA helps clustering methods discover subgroups within the data.

Although there is an improvement after PCA, the results of clustering are still non-ideal. For example, the homogeneity in clusters and dissimilarity between clusters is unclear. Most notably, k-medoid clusters created from dimensionally reduced data overlap. Therefore, observations in one cluster are not always dissimilar to observations in another. Hierarchal clustering yields better results when using dimensionally reduced data. The dendrogram shows a clear spilt into two clusters of different sizes and heights.

To further this assignment and potentially improve cluster results, I would experiment with dimension reduction and clustering methods. Dimension reduction shows improvements in cluster separation, dissimilarity, and density. Therefore, it is possible using fewer than four dimensions would improve the cluster results. Using three dimensions would constitute as a good solution as three principal components retain 80% of the variability within the data set. After reducing the data set to three dimensions, I would perform k-means clustering. This is because k-means clustering may yield more separated clusters than k-medoid clsuerting. K-means clusters differs from k-medoid clusters in their use of a centroid as the centre of a cluster instead of a medoid. It is possible the use of a centroid (which is the average of the members of the cluster) will change the central location of each cluster and thus alter the separation between clusters. The combination of three dimensions and k-means clustering may improve the evident overlap between clusters found in this assignment. However, a further loss of variability in the remaining dimensions might raise the question if the new clusters are still representative of the original data.

Despite issues with cluster separation and dissimilarity, the unsupervised methods still uncover meaning in the data. PCA biplots and loading vectors highlight spending associations between variables. Fresh, frozen and grocery products are grouped together, while spending on detergents and paper is not associated with spending on any products. Most significantly, after PCA, hierarchal clustering and k-medoid clustering reveal two clusters. This suggests overall there are two groups of client spending on wholesale products. However, the dissimilarity between these groups is debatable.

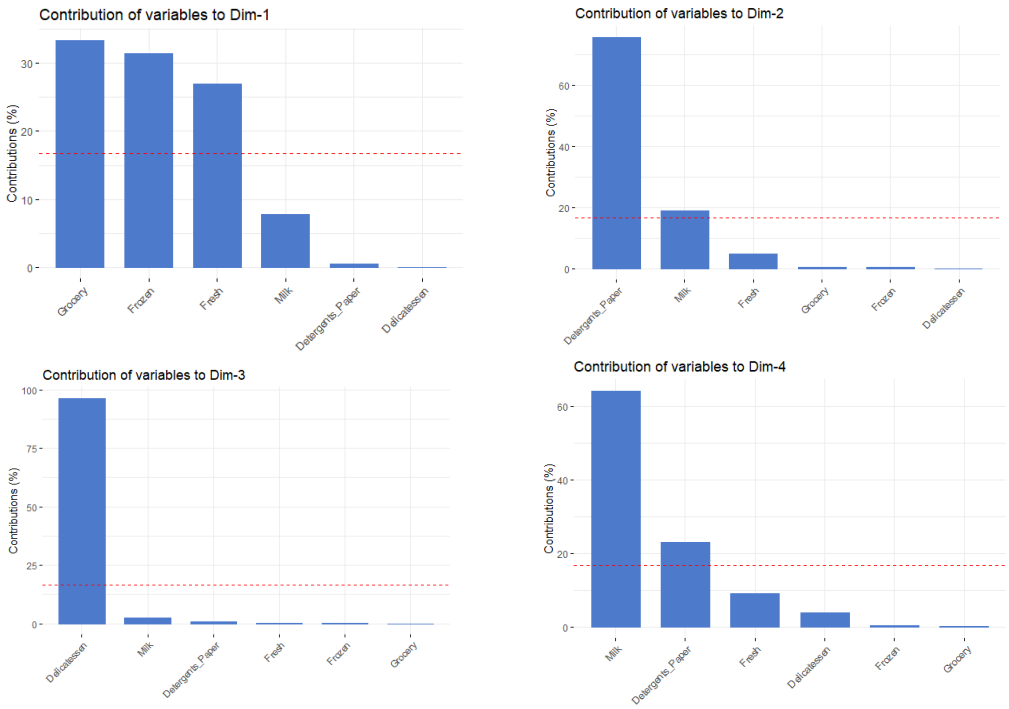
It is possible hierarchal clustering finds more dissimilarity between the two clusters than k-medoid clustering. This is because in hierarchical clustering, cluster one has significantly more observations in it than cluster two. Therefore, it is most probable the observations in the two groups are dissimilar. By contrast, the k-medoid clusters appear to be roughly equal in size, suggesting the observations may be similar in both clusters. The significant overlap between the k-medoid clusters also indicates the client spending may be similar in both clusters. It is important to note that multiple values outside of the k-medoid clusters suggest many client spending amounts are unpredictable.

In conclusion, the PCA results show there is at least one cluster category consisting of: fresh, grocery and frozen products. This pattern is also represented in the hierarchal and k-medoid clustering as the data separates into two clusters.

APPENDICES

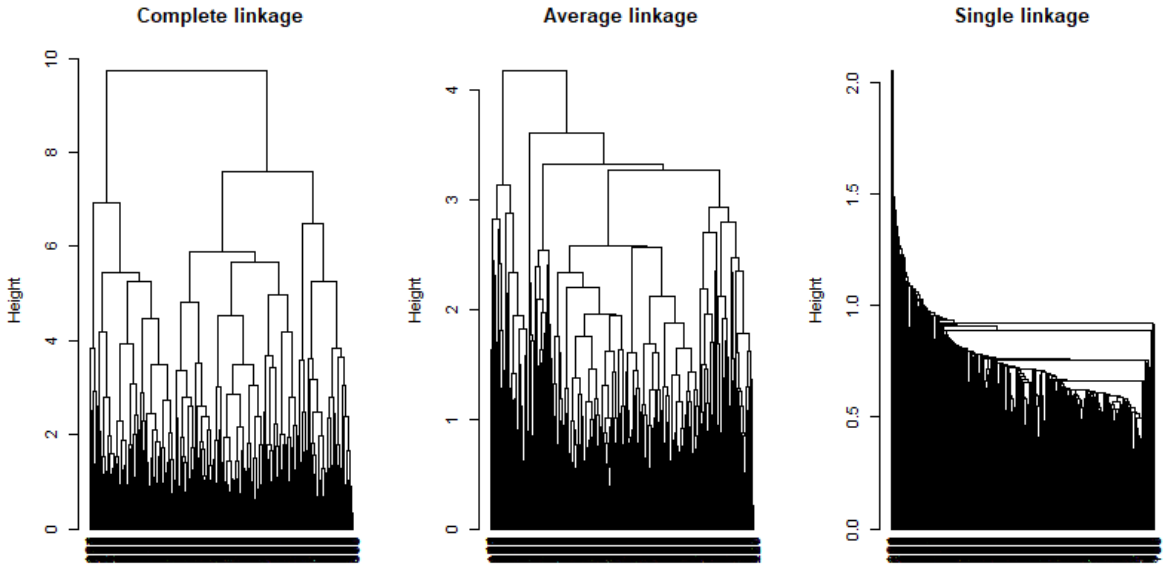
APPENDIX 1

Bar plots showing the contribution of variables to the four dimensions from PCA.



APPENDIX 2

Dendrograms for each linkage method using the dimensionally reduced data.



BIBLIOGRPAHY

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2020). An Introduction to Statistical Learning: with Applications in R. Springer Texts in Statistics. New York: Springer.

Kassambara, A. (2017), Practical Guide To Cluster Analysis in R: Unsupervised Machine Learning, STHDA

Katitas, A (2019) 'Getting Started with Multiple Imputation in R' https://data.library.virginia.edu/getting-started-with-multiple-imputation-in-r/ [Last accessed: 6/05/2022]

Rdocumentation. 'mice.impute.rf: Imputation by random forests' <https://www.rdocumentation.org/packages/mice/versions/3.14.0/topics/mice.impute.rf>[Last accessed: 6/05/2022]

Rdocumentation. 'mice.impute.midastouch: Predictive Mean Matching with Distance aided selection of donors'<https://www.rdocumentation.org/packages/midastouch/versions/1.3/topics/mice.impute.midastouchLast accessed: 6/05/2022]